VALUE IN HEALTH 15 (2012) 940–947

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Comparative Effectiveness Research/Health Technology Assessment

Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research

James E. Signorovitch, PhD[1,], Vanja Sikirica, PharmD, MPH[2], M. Haim Erder, PhD[2], Jipan Xie, MD, PhD[1], Mei Lu, MS[1], Paul S. Hodgkins, PhD, MSc[2], Keith A. Betts, PhD[1], Eric Q. Wu, PhD[1]*

[1]*Analysis Group, Inc., Boston, MA, USA;* [2]*Shire Development LLC, Wayne, PA, USA*

ABSTRACT

Objective: In the absence of head-to-head randomized trials, indirect comparisons of treatments across separate trials can be performed. However, these analyses may be biased by cross-trial differences in patient populations, sensitivity to modeling assumptions, and differences in the definitions of outcome measures. The objective of this study was to demonstrate how incorporating individual patient data (IPD) from trials of one treatment into indirect comparisons can address several limitations that arise in analyses based only on aggregate data. **Methods:** Matching-adjusted indirect comparisons (MAICs) use IPD from trials of one treatment to match baseline summary statistics reported from trials of another treatment. After matching, by using an approach similar to propensity score weighting, treatment outcomes are compared across balanced trial populations. This method is illustrated by reviewing published MAICs in different therapeutic areas. A novel analysis in attention deficit/hyperactivity disorder further demonstrates the applicability of the method. The strengths and limitations of MAICs are discussed in comparison to those of indirect comparisons that use only published aggregate data. **Results:** Example applications were selected to illustrate how indirect comparisons based only on aggregate data can be limited by cross-trial differences in patient populations, differences in the definitions of outcome measures, and sensitivity to modeling assumptions. The use of IPD and MAIC is shown to address these limitations in the selected examples by reducing or removing the observed cross-trial differences. An important assumption of MAIC, as in any comparison of nonrandomized treatment groups, is that there are no unobserved cross-trial differences that could confound the comparison of outcomes. **Conclusions:** Indirect treatment comparisons can be limited by cross-trial differences. By combining IPD with published aggregate data, MAIC can reduce observed cross-trial differences and provide decision makers with timely comparative evidence.

Keywords: comparative effectiveness, individual patient data, matching-adjusted indirect comparison.

Introduction

Health care decision makers face significant gaps between their needs for comparative effectiveness research (CER) and the limited availability of comparative data. The gap is particularly pronounced for new treatments, which are often integrated into treatment strategies and formulary policies without the benefits of randomized trials against all clinically or economically relevant alternatives. After the new treatment becomes available, observational studies based on registries or real-world data, or pragmatic trials, may be initiated. Such studies, however, will not provide reliable comparative evidence until sufficient outcomes data have accumulated. This delay in comparative evidence for new treatments reduces the value of CER for improving the decisions of physicians, payers, and patients.

An increasingly used approach for timely CER is the comparison of treatment outcomes across separate randomized trials. Detailed reviews of methodologies for such indirect comparisons have been published [1,2], and guidelines have been developed for researchers and decision makers [3–10]. By combining trials with overlapping comparator groups, multiple direct and indirect comparisons can be combined into a network meta-analysis that summarizes comparative evidence for all treatments in a therapeutic area. Although based on randomized trials, indirect comparisons and network meta-analyses involve comparisons of nonrandomized treatment groups and are akin to observational studies and subject to important limitations.

In particular, cross-trial differences in patients' baseline characteristics or differences in outcome definitions can bias indirect comparisons [1,2,11]. Although meta-regressions can adjust for cross-trial differences in baseline characteristics at an aggregate level, they cannot adjust for large numbers of baseline differences and may be subject to ecological bias [12,13]. A key assumption of indirect comparisons and network

We declare the following potential conflicts of interest: Vanja Sikirica, M. Haim Erder, and Paul S. Hodgkins are employees of Shire *Pharmaceuticals, Inc.*, and hold stock options in Shire *Pharmaceuticals, Inc.* James E. Signorovitch, Jipan Xie, Mei Lu, Keith A. Betts, and Eric Q. Wu are employees of Analysis Group, Inc., a consultant for Shire *Pharmaceuticals, Inc.*

* *Address correspondence to:* James E. Signorovitch, Analysis Group, Inc., 111 Huntington Avenue, 10th floor, Boston, MA 02199.

E-mail: jsignorovitch@analysisgroup.com.

1098-3015/$36.00 – see front matter Copyright © 2012, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

http://dx.doi.org/10.1016/j.jval.2012.05.004

meta-analyses is that cross-trial differences can be mitigated by measuring treatment effects relative to a common comparator (e.g., placebo). However, different modeling assumptions, embodied in the choice of a relative effect measure (e.g., relative risk, odds ratio, or risk difference), can lead to conflicting conclusions about comparative effectiveness. These limitations are difficult to address by using only published aggregate data, especially when only small numbers of trials are available.

In this article, we show that the use of individual patient data (IPD) from clinical trials for one treatment—but not necessarily all treatments—can address these limitations and that useful IPD are more readily available than currently appreciated. If IPD were available from all trials of interest, potential biases stemming from cross-trial differences could be mitigated by regression adjustment [14–18] or propensity scores [19,20]. Although IPD are seldom available for all trials, researchers engaged in CER can often access IPD for some trials. In particular, when CER is conducted or funded by a clinical trial sponsor, IPD could be available from the sponsor's trials. A recently developed statistical method—*matching-adjusted indirect comparison (MAIC)*—can combine IPD for some treatments with published summary data for comparator treatments. Through examples, we show how the incorporation of IPD into indirect comparisons via MAIC can 1) adjust for cross-trial differences in baseline characteristics, 2) reduce sensitivity to effect measures, 3) resolve differences in study outcome definitions, and 4) allow the comparison of clinically relevant dosages.

Methods

The MAIC approach has been published previously and is briefly reviewed below [21]. The approach can be applied in three steps.

Clinical trial selection

As in any data synthesis, a systematic review should be conducted to identify clinical trials for the treatments to be compared. Characteristics of the selected trials should then be carefully compared, including the study design (e.g., randomized or open-label trial), inclusion/exclusion criteria, baseline characteristics, outcome assessments (e.g., definitions of outcomes, schedule of assessments), and statistical methods (e.g., handling of early dropouts, baseline adjustment). Cross-trial differences in these features can be sources of heterogeneity in any meta-analysis. In indirect comparisons and network meta-analyses, these differences can also be sources of bias. The availability of IPD, which can provide opportunities to remove or reduce observed cross-trial differences, should be assessed for each trial.

Identification of outcome measures

A meaningful cross-trial comparison should focus on comparably defined outcome measures that are available in the included trials. The precise definition of the study outcomes, the schedule of assessments, the clinical relevance of different dosages, and the statistical methods used to summarize effects should all be considered. IPD should be reanalyzed to match the outcome definitions used in the published trial data as much as possible before making an indirect comparison. If outcome definitions cannot be matched exactly, sensitivity analyses should be considered.

Matching trial populations

In trials with IPD, patients who could not have enrolled in the published comparator trials (e.g., because of stricter inclusion/exclusion criteria) should be excluded from the indirect comparison analysis. Even after matching inclusion/exclusion criteria across trials, important cross-trial differences in patients' baseline characteristics can remain. To adjust for these differences by using MAIC, patients in trials with IPD are weighted such that their weighted mean baseline characteristics match those reported for the trials without IPD. This approach is a form of propensity score weighting in which patients in one treatment group (in this case the trial with IPD) are weighted by their inverse odds of being in that group versus the other treatment group (in this case the trial with only published aggregate data). The propensity score model can be estimated by using the generalized method of moments based on the aggregate data and IPD. Other baseline summary statistics such as medians and standard deviations can also be matched when available. Outcomes from common comparator arms (e.g., placebo) can be used to validate the matching process. After matching, continuous, binary, or time-to-event outcomes can be compared across balanced trial populations by using weighted statistical tests that incorporate the same weights developed in the matching process (e.g., using weighted *t* tests, weighted χ^2 tests, or Kaplan-Meier tests). Weighted statistical models (e.g., analysis of covariance) can also be used to ensure that similar methods are applied to all trials. Limitations of the MAIC approach are described in the Discussion section.

Example applications

To illustrate how the use of IPD with MAIC can address the limitations that arise for indirect comparisons without IPD, four example applications are presented.

Indirect comparisons with IPD can resolve significant differences in key baseline characteristics: vildagliptin versus sitagliptin in Japanese patients with type II diabetes mellitus [22]

Vildagliptin and sitagliptin are two treatments for type II diabetes that were recently approved for use in Japan. Both have been associated with better glycemic control compared with placebo or voglibose in randomized trials [23–29]. A systematic literature review of clinical trials in Japanese patients identified two vildagliptin [25,26] and two sitagliptin [23,27] trials. The common comparators included placebo and voglibose. An indirect comparison of aggregate data suggested that vildagliptin was associated with a significantly greater absolute decrease in mean % glycosylated hemoglobin A_{1c} (Hb A_{1c}) versus sitagliptin (difference = −0.17; 95% confidence interval [CI]: −0.33 to −0.01; $P = 0.024$) in the voglibose-controlled trials but that the difference was not statistically significant in the placebo-controlled trials (difference = −0.20; 95% CI: −0.45 to 0.05; $P = 0.133$). Significant cross-trial baseline differences in mean Hb A_{1c}, however, call into question the validity of the comparison based only on published aggregate data. Patients in the vildagliptin trials had significantly lower mean Hb A_{1c} at baseline than did patients in the sitagliptin trials. Higher baseline Hb A_{1c} has been associated with greater postbaseline Hb A_{1c} reduction in meta-analyses of oral antihyperglycemics [30]. Although meta-regression can adjust for baseline differences in aggregate trial data, it can be unreliable with data from only four trials.

By using IPD from the vildagliptin trials, patients were selected on the basis of inclusion/exclusion criteria specified in the sitagliptin trials and were reweighted to match exactly the baseline characteristics reported for the sitagliptin trials, including the baseline mean Hb A_{1c} as well as age, sex, body mass index, fasting plasma glucose (FPG), and diabetes duration. After matching, vildagliptin was associated with a significantly greater decrease in Hb A_{1c} compared with sitagliptin (Table 1). Compared with the indirect comparison based on only aggregate data, the treatment difference between vildagliptin and sitagliptin increased after MAIC, consistent with the expected effect of adjusting for baseline Hb A_{1c} differences. The use of IPD and MAIC in this example provided a more reliable comparison than did aggregate data alone,

Table 1 – Selected baseline characteristics and outcomes before and after matching-vildagliptin vs. sitagliptin in the treatment of japanese patients with type 2 diabetes mellitus[22].

Characteristics or outcomes	Before matching		After matching	
	Vildagliptin trials[25,26]	Sitagliptin trials[23,27]	Vildagliptin trials[25,26]	Sitagliptin trials[23,27]
Baseline characteristics				
Voglibose-controlled trials of vildagliptin 50mg bid (N=363) or sitagliptin 50mg od (N=319)				
Age [year; mean (SD)]	59.1 (9.9)*	60.7 (10.0)	60.7 (10.0)	60.7 (10.0)
HbA1c [%; mean (SD)]	7.6 (0.9)†	7.8 (0.9)	7.8 (0.9)	7.8 (0.9)
FPG [mg/dL; mean (SD)]	162.4 (31.3)‡	149.4 (32.4)	149.4 (32.4)	149.4 (32.4)
Placebo-controlled trials of vildagliptin 50mg bid (N=148) or sitagliptin 50mg od (N=145)				
Age [year; mean (SD)]	59.6 (8.3)	60.2 (8.7)	60.2 (8.7)	60.2 (8.7)
HbA1c [%; mean (SD)]	7.4 (0.8)†	7.7 (0.9)	7.7 (0.9)	7.7 (0.9)
FPG [mg/dL; mean (SD)]	161.4 (31.5)†	150.4 (32.1)	150.4 (32.1)	150.4 (32.1)
Outcome: HbA1c during the 12-month outcomes period [%, mean (95% CI)]				
Voglibose-controlled trials				
Vildagliptin 50mg bid vs. sitagliptin 50mg od	−0.57 (−0.69, −0.45)	−0.40 (−0.50, −0.30)	−0.60 (−0.72, −0.48)	−0.40 (−0.50, −0.30)
Difference	−0.17 (−0.33, −0.01)*		−0.20 (−0.36, −0.04)†	
Placebo-controlled trials				
Vildagliptin 50mg bid vs. sitagliptin 50mg od	−1.19 (−1.39, −0.99)	−0.99 (−1.17, −0.81)	−1.55 (−1.80, −1.30)	−0.99 (−1.17, −0.81)
Difference	−0.20 (−0.45, 0.05)		−0.56 (−0.87, −0.25)‡	
Pooled				
Difference of vildagliptin 50mg bid vs. sitagliptin 50mg od	−0.18 (−0.30, −0.06)†		−0.28 (−0.42, −0.14)‡	

Chi-squared tests for binary baseline characteristics and t-tests for continuous baseline characteristics were used before matching; for comparison of HbA1c during the 12-month outcome period, unweighted ANCOVA before matching and weighted ANCOVA after matching were used
ANCOVA, analysis of covariance; bid, twice daily; HbA1c, FPG, fasting plasma glucose; glycosylated haemoglobin; od, once daily.
* $p < 0.05$.
† $p < 0.01$, and
‡ $p < 0.001$ for comparisons of vildagliptin vs. sitagliptin trials.

despite the small number of trials available for the Japanese population.

Indirect comparisons with IPD can reduce sensitivity to the effect measure: adalimumab versus etanercept in the treatment of psoriasis [21]

Adalimumab (ADA) and etanercept (ETN) are the most commonly used Food and Drug Administration –approved anti–tumor necrosis factors for the treatment of psoriasis. Both have demonstrated efficacy relative to placebo for the reduction of psoriasis skin lesions in randomized trials [31–35]. A systematic literature review identified two ADA trials with IPD [32,35] and one ETN trial [33] with summary results as having sufficiently comparable designs. The efficacy outcome considered here, an improvement of 90% or more in the Psoriasis Area and Severity Index (PASI 90) score from baseline to week 12, was higher in the ADA trials than in the ETN trial (37.0% vs. 22.0%). The placebo-arm response rates, however, were also higher in the ADA trials (1.4% vs. 0.6%) (Table 2). Indirect comparisons of these PASI 90 rates based only on aggregate data can give results in opposing directions, depending on whether the rate difference or odds ratio is used as the relative effect measure. When using the rate difference, ADA is associated with an absolute 14.2% greater rate of PASI 90 compared with ETN; when using the odds ratio, ADA is associated with 13% lower odds of PASI 90 compared with ETN, with an odds ratio of 0.87. Such sensitivity to the effect measure arises from differences in placebo-arm response rates and is difficult to resolve by using only published aggregate data. Although the odds ratio is often preferred for indirect comparisons, and has desirable statistical properties, there is no evidence that it provides a better model for cross-trial adjustment than the risk difference in this application to a small number of trials.

After using MAIC with IPD for the ADA versus placebo trials to remove observed differences in mean baseline characteristics compared with the ETN versus placebo trial, the difference in placebo-arm response rates was reduced to 0.9% versus 0.6%, for the ADA and ETN trials respectively. Comparison of the matched populations resulted in a consistently higher rate of PASI 90 with ADA versus ETN (rate difference = 14.8%; 95% CI 6.8%–22.8%; $P < 0.001$), with a corresponding odds ratio of 1.34 indicating a 34% increase in the odds of PASI 90 with ADA versus ETN. Thus, using IPD to match trial populations reduced the sensitivity to the effect measure that would have limited an indirect comparison based only on aggregate data.

Indirect comparisons with IPD can resolve differences in outcome definitions: nilotinib versus dasatinib in newly diagnosed chronic myelogenous leukemia chronic phase [36]

Nilotinib and dasatinib, both recently approved by the Food and Drug Administration for the treatment of newly diagnosed chronic myelogenous leukemia (CML) In the chronic phase, have superior efficacy compared with the previous standard of care, imatinib [37,38]. There are only two published trials of nilotinib or dasatinib for CML chronic phase (ENESTnd and DASISION), each compared with imatinib [39,40]. Although major molecular response (MMR) was reported for both trials, each trial used a different definition for this endpoint. The ENESTnd trial reported MMR *at* 12 months, while DASISION reported MMR *by* 12 months. In addition, MMR was assessed at different times, which is important for assessing achievement by 12 months because some responses can be transient. In ENESTnd,

Table 2 – Selected baseline characteristics and outcomes before and after matching – ADA vs. ETN in the Treatment of Psoriasis[21].

Characteristics or outcomes	Before matching		After matching	
	ADA trials[32,35] (N=1052)	ETN trial[33] (N=330)	ADA trials[32,35] (N=1052)	ETN trial[33] (N=330)
Baseline characteristics				
Age [year; mean (SD)]	43.9 (13.2)†	45.2 (11.6)	45.2 (11.6)	45.2 (11.6)
Psoriatic arthritis (%)	26.9†	22.0	22.0	22.0
Prior systemic or phototherapy (%)	62.2‡	76.0	76.0	76.0
Affected body surface area [mean% (SD)]	25.9 (15.0)‡	29.3 (19.3)	29.3 (19.3)	29.3 (19.3)
Outcome: week 12 PASI response ≥90% (%)				
PBO in ADA trials (n=347) vs. PBO in the ETN trial (n=166)	1.4	0.6	0.9	0.6
ADA (n=678) vs. ETN (n=164)	37.0	22.0	37.1	22.0
Difference*			14.8 (6.8, 22.8)§	

ADA, adalimumab; ETN, etanercept; PASI, Psoriasis Area and Severity Index; PBO, placebo.
* Difference-in-difference of response rates: (ADA - PBO in ADA trials) - (ETN - PBO in the ETN trial).
† $p < 0.05$.
‡ $p < 0.01$, and
§ $p < 0.001$ for comparisons of ADA vs. ETN trials.
Chi-squared tests for binary baseline characteristics and t-tests for continuous baseline characteristics were used before matching; a weighted chi-squared test was used for week 12 PASI response ≥90% after matching.

MMR was assessed at months 1, 2, 3, 6, 9, and 12, while in DASISION, it was assessed within 6 weeks after randomization and then at months 3, 6, 9, and 12. Because of the different definitions and assessment schedules, comparison of MMR outcomes in an indirect comparison using aggregate data could be biased. Based solely on the aggregate data, the treatments appeared to have similar efficacy; the MMR *at* 12 months was 44% for nilotinib, and the MMR *by* 12 months was 46% for dasatinib.

After using IPD from ENESTnd to compute MMR *by* 12 months in that trial (using only months 3, 6, 9, and 12), however, treatment with nilotinib was associated with a significantly higher rate of MMR *by* month 12 compared with dasatinib (56.8% vs. 45.9%; rate difference = 10.8%; 95% CI 2.2%–19.4%; P = 0.014; Table 3) after adjusting for multiple baseline characteristics by using MAIC, including age, sex, CML duration, performance status, and laboratory markers associated with CML severity. No difference was seen in the rate of MMR between the imatinib arms after matching (rate difference = −1.2%; 95% CI −9.04% to 6.64%; P = 0.77). This example shows the importance of identifying and reconciling differences in outcome definitions in indirect comparisons.

Indirect comparisons using IPD can compare clinically relevant dosages: guanfacine extended release versus atomoxetine in children and adolescents with attention deficit/hyperactivity disorder

Guanfacine extended release (GXR) and atomoxetine (ATX) are two commonly used Food and Drug Administration–approved

Table 3 – Selected baseline characteristics and outcomes before and after matching– nilotinib vs. dasatinib in the treatment newly diagnosed CML-CP patients[36].

Characteristics or outcomes	Before matching		After matching	
	Nilotinib trial[39] (N=273)	Dasatinib trial[40] (N=259)	Nilotinib trial [39] (N=273)	Dasatinib trial[40] (N=259)
Baseline characteristics				
Age (year)				
Median	47	46	46	46
≥65 (%)	11	8	8	8
Eastern Cooperative Oncology Group Performance Status (%)				
0	88	82	82	82
1	12	18	18	18
Platelet count [median ($\times 10^{-3}$/mm^3)]	419	448	448	448
Outcome: difference in major molecular response (MMR) by 12 months [% (95% CI)]				
imatinib 400mg od in nilotinib trials vs. imatinib in dasatinib trials	NR		−1.2 (−9.0, 6.6)	
nilotinib 300mg bid vs. dasatinib 100mg od	9.4 (1.0, 17.8)*		10.8 (2.2, 19.4)*	

For the comparison of MMR by 12 month, a chi-squared test was used before matching and a weighted chi-squared test was used after matching.
Bid, twice daily, od, once daily, NR, not reported.
* $p < 0.05$ for comparisons of nilotinib vs. dasatinib trials.

nonstimulants for the treatment of attention deficit/hyperactivity disorder (ADHD). A systematic literature review identified two trials with IPD for GXR [41,42] and four trials with summary results for ATX [43–46]. A key difference between these trials was the assignment of dosages: the GXR trial randomized patients to fixed doses, regardless of body weight, whereas the ATX trial randomized patients to weight-based doses of 0.5 to 2.0 mg/kg/d. Within the GXR trial, weight-based doses ranged from 0.01 to 0.17 mg/kg/d. Efficacy comparisons at doses considered to be therapeutic alternatives are necessary to provide meaningful comparative evidence to decision makers, but indirect comparisons are difficult to conduct by using aggregate when outcomes are reported for fixed versus weight-based dosing. By using IPD data from the GXR trials, patients were selected if they met the inclusion/exclusion criteria of the ATX trials and were expected to receive target doses of 0.09 to 0.12 mg/kg/d [47]. The change in ADHD Rating Scale, Version IV (ADHD-RS-IV) total score from baseline to end point was compared between this GXR-treated population and ATX treatment at a target dose of 1.2 mg/kg/d [48]. Lower doses of GXR (0.046–0.075 and 0.075–0.090 mg/kg/d) were also considered as sensitivity analyses. In each of the dosage groups, GXR-treated patients were reweighted to match the mean baseline characteristics of the ATX populations, including age, sex, weight, ADHD subtype, baseline ADHD-RS-IV scores, and average placebo outcome. Moreover, because only one ATX trial included a 1.2 mg/kg/d arm, a sensitivity analysis including all ATX arms with a maximum allowed dose of 1.2 mg/kg/d or more was conducted, pooling data from all four ATX trials.

Before matching, patients treated with GXR had a greater percentage of the combined ADHD subtype versus those treated with 1.2 mg/kg/d or more of ATX (73.2% vs. 69.3%). In unadjusted indirect comparisons of these data, the ADHD-RS-IV score change from baseline was higher in each weight-based dosing group for patients treated with GXR versus those treated with ATX 1.2 or 1.2 mg/kg/d or more, respectively. The placebo arm in the GXR trials, however, also showed greater ADHD-RS-IV score reduction than in the ATX trials (Table 4).

After matching, baseline characteristics and placebo-arm responses were identical between the two populations (Table 4). In the base case, compared with ATX 1.2 mg/kg/d, GXR 0.09 to 0.12 mg/kg/d was associated with a significantly greater reduction in the mean ADHD-RS-IV total score (mean difference = −7.0; 95% CI −11.3 to −2.7; $P < 0.01$). In the sensitivity analysis, GXR 0.075 to 0.090 mg/kg/d was also associated with significantly greater reduction in the mean ADHD-RS-IV total score (mean difference = −6.0; 95% CI −11.3 to −0.7; $P < 0.05$). This was also true for GXR 0.09 to 0.12 and 0.046 to 0.075 mg/kg/d compared with ATX 1.2 mg/kg/d or higher doses.

By using IPD, it was possible to select specific GXR patients with doses that are considered therapeutic alternatives to ATX doses that have been investigated in randomized trials. Matching further adjusted for observed baseline characteristics and placebo efficacy across trials to ensure a comparison across balanced trial populations.

Discussion

Through several example applications, this article has demonstrated how the use of IPD and MAIC can provide timely and reliable comparisons between treatments when indirect comparisons based only on published aggregate data are limited by cross-trial differences. Each of the example applications has illustrated a different benefit of using available IPD and MAIC. First, IPD and MAIC removed mean differences in a key baseline characteristic (Hb A_{1c}) in diabetes trials. Second, the use of IPD and MAIC reduced the differences in placebo-arm response rates among psoriasis trials, rendering the indirect comparison less sensitive to the choice of a relative effect measure (risk difference or odds ratio). Third, using IPD resolved differences in study protocols and end-point definitions for MMR in CML trials. Finally, the use of IPD and MAIC allowed comparison of clinically relevant dosages with adjustment for baseline symptom scores in ADHD trials.

All these examples have included relatively small numbers of trials. Such settings often arise as new treatments become available. In some cases, as in the CML example, only one trial is available for each treatment. These settings are particularly challenging for methods based on aggregate data because having a small number of trials may increase the chances that trials of one treatment have significantly different characteristics or designs than do trials of another treatment. The underlying model for indirect comparisons based on aggregate data assumes that observed trials represent an exchangeable random sample from an underlying distribution of all possible trials of the studied treatments [1,2,49,50]. But even if this assumption holds true, a random sample of a small number of trials can risk substantial imbalance when trial-level heterogeneity is high. To the extent that observed factors such as inclusion/exclusion criteria, patient baseline characteristics, and outcome definitions can explain cross-trial heterogeneity, the use of IPD and MAIC can improve the reliability of the indirect comparisons compared with using aggregate data only.

Many researchers can access IPD for some treatments of interest in a comparative analysis, but accessing IPD for all treatments is often not possible. Without an appropriate methodology for combining IPD with publicly available summary data, IPD has remained a largely untapped resource for CER. By utilizing IPD available only for one comparator treatment, MAICs could substantially increase the availability and reliability of comparative evidence. The approach could be especially helpful for new treatments with small number of trials, for which many treatment and reimbursement decisions are currently based only on informal (naive) indirect comparisons across trials or on anchor-based indirect comparisons subject to the limitations described above. Payers or health technology assessment authorities responsible for formulary policy, who routinely request indirect comparisons and economic models from manufacturers, can consider MAIC as a way to fill evidence gaps and increase the reliability of the comparative evidence on which they base their decisions. A common goal of payers in synthesizing comparative evidence is to ensure that all relevant and available data are utilized. When IPD are available, as in manufacturer submissions, these data can be utilized with MAIC and, as shown the examples above, can be highly relevant.

Limitations

Indirect comparisons, like any comparison of nonrandomized treatment groups, can be biased by both observed and unobserved cross-trial differences. Although the use of IPD and MAIC can remove or reduce observed cross-trial differences, unobserved differences may result in residual confounding. Comparing placebo-arm (or other common comparator arm) outcomes across trials provides an opportunity to detect some residual confounding. Greater differences in placebo-arm outcomes indicate a greater chance of important unobserved differences between trials and greater sensitivity to the choice of a relative effect measure (e.g., odds ratio vs. risk difference). Even when placebo-arm outcomes are well balanced between trials, however, an indirect comparison may be biased by unobserved factors that impact treatment- but not placebo-arm outcomes (e.g., baseline factors impacting adherence). Only a well-controlled head-to-head randomized trial can avoid unobserved confounding.

Practical limitations of IPD and MAIC are important. First, it is not always possible to match outcome definitions or inclusion/exclusion criteria by using IPD. For example, if a published trial

Table 4 – Selected baseline characteristics and outcomes before and after matching – GXR vs. ATX in the treatment of ADHD.

Characteristics or outcomes	Before matching		After matching	
	GXR trials[41,42] (N=403)	ATX trials[43,44,45,46] (N=873)	GXR trials[41,42] (N=403)	ATX trials[43,44,45,46] (N=873)
Baseline characteristics	10.4 (2.6)	10.2 (2.0)	10.2 (2.0)	10.2 (2.0)
Age [years; mean(SD)]	74.2	73.4	73.4	73.4
Male (%)				
ADHD subtype (%)	24.3	28.6	28.6	28.6
Inattentive	2.5	2.1*	2.1	2.1*
Hyperactive-impulsive	73.2	69.3	69.3	69.3
Combined				
ADHD-RS-IV scores [mean (SD)]				
Total score	39.4 (8.5)	39.5 (8.7)	39.5 (8.7)	39.5 (8.7)
Hyperactivity/impulsivity subscale score	17.4 (6.5)	17.6 (6.6)	17.6 (6.6)	17.6 (6.6)
Inattentive subscale score	22.0 (4.1)	21.9 (3.9)	21.9 (3.9)	21.9 (3.9)
Outcome: ADHD-RS-IV Total scores change from baseline [Mean (95% CI)]				
Placebo in GXR trials (n=136) vs. placebo in the ATX trial† (n=83)	−10.6 (−12.8, −8.4)	−5.8 (−8.2, −3.4)	−5.8 (−8.2, −3.4)	−5.8 (−8.2, −3.4)
Difference	−4.8 (−8.0, −1.6)§		0	
GXR 0.046-0.075 mg/kg/day (n=147) vs. ATX 1.2 mg/kg/day† (n=84)	−17.9 (−20.1, −15.7)	−13.6 (−16.5, −10.7);	−17.3 (−19.8, −14.8)	−13.6 (−16.5, −10.7)
Difference	−4.3 (−7.9, −0.6)§		−3.7 (−7.6, 0.2)	
GXR 0.075-0.090 mg/kg/day (n=46) vs. ATX 1.2 mg/kg/day† (n=84)	−18.5 (−22.6, −14.4)	−13.6 (−16.5, −10.7)	−19.6 (−23.9, −15.3)	−13.6 (−16.5, −10.7)
Difference	−4.9 (−10.0, 0.2)		−6.0 (−11.3, −0.7)‡	
GXR 0.09-0.12 mg/kg/day (n=82) vs. ATX 1.2 mg/kg/day† (n=84)	−22.8 (−25.3, −20.3)	−13.6 (−16.5, −10.7)	−20.6 (−23.7, −17.5)	−13.6 (−16.5, −10.7)
Difference	−9.2 (−13.1, −5.3)‡		−7.0 (−11.3, −2.7)§	
Placebo in GXR trials (n=136) vs. placebo in ATX trials (n=348)	−10.6 (−12.8, −8.4)	−5.8 (−7.0, −4.6)	−5.8 (−7.0, −4.6)	−5.8 (−7.0, −4.6)
Difference	−4.8 (−7.3, −2.3)§		0	
GXR 0.046-0.075 mg/kg/day (n=147) vs. ATX ≥1.2 mg/kg/day (n=506)	−17.9 (−20.1, −15.7)	−14.6 (−15.8, −13.4)	−18.9 (−21.3, −16.5)	−14.6 (−15.8, −13.4)
Difference	−3.3 (−5.7, −0.8)§		−4.3 (−6.8, −1.8)§	
GXR 0.075-0.090 mg/kg/day (n=46) vs. ATX ≥1.2 mg/kg/day (n=506)	−18.5 (−22.6, −14.4)	−14.6 (−15.8, −13.4)	−17.7 (−21.8, −13.6)	−14.6 (−15.8, −13.4)
Difference	−3.9 (−8.2, 0.4)		−3.1 (−7.2, 1.0)	
GXR 0.09-0.12 mg/kg/day (n=82) vs. ATX ≥1.2 mg/kg/day (n=506)	−22.8 (−25.3, −20.3)	−14.6 (−15.8, −13.4)	−22.2 (−24.7, −19.7)	−14.6 (−15.8, −13.4)
Difference	−8.2 (−11.0, −5.5)‖		−7.6 (−10.3, −4.9)‖	

GXR=guanfacine extended-release (Intuniv) ; ATX = atomoxetine (Strattera).

* Unspecified and hyperactive-impulsive categories were combined for ATX trials.

† Only one ATX trial [46] included ATX arm 1.2 mg/kg/day; the corresponding placebo arm in the same trial was used for this analysis.

‡ $p < 0.05$.

§ $p < 0.01$, and

‖ $p < 0.001$ for comparisons of GXR vs ATX trials. Chi-squared tests for binary baseline characteristics and t-tests for continuous baseline characteristics were used before matching; for comparison of ADHD-RS-IV total scores change from baseline, t-tests were used before matching and weighted t-tests were used after matching.

reports response by month 12, based on assessments every 3 months, and the trial with IPD has assessments only every 6 months, a fair comparison may be impossible. Likewise, if the published trial enrolled patients of all ages, but IPD are available only for patients younger than 65 years, it will be impossible to fully match the age distribution of the published trial. Second, unlike traditional propensity score weighting, the availability of only aggregate data for some trials in an MAIC prevents the use of existing methods for checking the fit and calibration of the propensity score model. Despite this limitation, balance in the average baseline characteristics can be verified in MAIC, as in traditional propensity score matching. Third, the ability to adjust for multiple baseline factors depends on having a sufficient number of patients in trials with IPD. In general, matching larger numbers of average baseline characteristics and adjusting for greater cross-trial baseline differences will require more extreme weights and will reduce the effective sample size. In practice, this is less of a limitation than faced by meta-regression, which requires the number of trials, rather than the number of patients, to exceed the number of baseline characteristics used for adjustment. However, in extreme cases, as when there are more baseline characteristics than patients, it should be noted that MAIC cannot be used to match all baseline characteristics. Similarly, if there are extreme differences in multiple baseline characteristics (e.g., in some trials, 90% are treatment naive and 5% are aged >65 years, yet in other trials, only 10% are treatment naive and 60% are aged >65 years), it may be impossible to balance all baseline characteristics by using MAIC, or any other method, if the trials populations truly have limited overlap.

MAIC may also be applied to single-arm trials, or trials without a common comparator. In these situations, which are common for

surgical interventions, rare diseases, and oncology trials for patients with poor prognosis, the use of IPD may be the only way to adjust for cross-trial differences and should be preferred over naive unadjusted comparisons. The absence of a common comparator arm, however, should be noted as an important limitation, because validation of the matching or the use of relative effect measures will not be possible.

Finally, when using IPD and MAIC, there may be trials identified in a systematic review, either with IPD or with only aggregate data available, that cannot be included in the analysis because of irreconcilable differences in design or patient characteristics. Although the consideration of all available evidence is an important guiding principle for any evidence synthesis, there is a trade-off between including more evidence and reducing heterogeneity. For indirect comparisons, heterogeneity across trials can result in confounding and bias. Avoidance of bias should be a primary goal of indirect comparisons, and the need to reduce bias should be considered as thoroughly as the need to include all trials. In cases where the inclusion or exclusion of a trial is questionable, sensitivity analyses should be conducted with and without it.

Conclusion

While randomized trials are the gold standard for CER, they are not always available for clinically and economically important treatment comparisons, especially when novel therapies become available. Many researchers engaged in CER can access IPD for certain trials, but these data have not been widely used for indirect comparisons. The examples presented in this article have shown that by using IPD from one treatment and published aggregate data from another treatment, MAICs can provide greater adjustment for observed cross-trial differences compared with indirect comparisons based only on published aggregate data. In this way, the use of IPD and MAIC can provide decision makers with timely and reliable comparative evidence.

Source of financial support: Shire Development LLC, provided funding to Analysis Group, Inc., to perform this study. Although the sponsor was involved in the design, collection, analysis, interpretation, and fact checking of information, the content of this manuscript, the ultimate interpretation, and the decision to submit it for publication was not contingent on the sponsor's approval.

REFERENCES

[1] Sutton A, Ades AE, Cooper N, et al. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 2008;26:753–67.
[2] Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;21:2313–24.
[3] Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices, Part 2. Value Health 2011;14:429–37.
[4] Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices, Part 1. Value Health 2011;14:417–28.
[5] Wells GA, Sultan SA, Chen L, et al. Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis. Ottawa, Ontario, Canada: Canadian Agency for Drugs and Technologies in Health, 2009.
[6] NICE Guide to the Methods of Technology Appraisal. Available from http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf. [Accessed May 18, 2012].
[7] Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 1: introduction to evidence synthesis for decision making. 2011. Available from http://www.nicedsu.org.uk. [Accessed May 18, 2012].
[8] Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. 2011; last updated August 2011. Available from http://www.nicedsu.org.uk. [Accessed May 18, 2012].
[9] Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 3: heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011. Available from http://www.nicedsu.org.uk. [Accessed May 18, 2012].
[10] Dias S, Welton NJ, Sutton AJ., et al. NICE DSU Technical Support Document 4: inconsistency in networks of evidence based on randomised controlled trials. 2011. Available from http://www.nicedsu.org.uk. [Accessed May 18, 2012].
[11] Chou R, Fu R, Hoyt Human L, Korthuis P. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. Lancet 2006;368:1503–15.
[12] Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989;18:269–74.
[13] Berlin JA, Santanna J, Schmid CH, et al. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med 2002;21:371–87.
[14] Pignon JP, Arriagada R, Ihde DC, et al. A meta-analysis of thoracic radiotherapy for small-cell lung cancer. N Engl J Med 1992;327:1618–24.
[15] Homocysteine Lowering Trialists' Collaboration. Lowering blood homocysteine with folic acid based supplements: meta-analysis of randomised trials. BMJ 1998;316:894–8.
[16] Non-small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. BMJ 1995;311:899–909.
[17] Moore RA, McQuay HJ. Single-patient data metaanalysis of 3453 postoperative patients: oral tramadol versus placebo, codeine and combination analgesics. Pain 1997;69:287–94.
[18] Turner RM, Omar RZ, Yang M, et al. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 2000;19:3417–32.
[19] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
[20] Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv Outcomes Res Method 2001;2:259–78.
[21] Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. Pharmacoeconomics 2010;28:935–45.
[22] Signorovitch JE, Wu EQ, Swallow E, et al. Comparative efficacy of vildagliptin and sitagliptin in Japanese patients with type 2 diabetes mellitus: a matching-adjusted indirect comparison of randomized trials. Clin Drug Investig 2011;31:665–74.
[23] Iwamoto Y, Tajima N, Kadowaki T, et al. Efficacy and safety of sitagliptin monotherapy compared with voglibose in Japanese patients with type 2 diabetes: a randomized, double-blind trial. Diabetes Obesity Metab 2010;12:613–22.
[24] Iwamoto Y, Taniguchi T, Nonaka K, et al. Dose-ranging efficacy of sitagliptin, a dipeptidyl peptidase-4 inhibitor, in Japanese patients with type 2 diabetes mellitus. Endocr J 2010;57:383–94.
[25] Iwamoto Y, Kashiwagi A, Yamada N, et al. Efficacy and safety of vildagliptin and voglibose in Japanese patients with type 2 diabetes: a 12-week, randomized, double-blind, active-controlled study. Diabetes Obesity Metab 2010;12:700–8.
[26] Kikuchi M, Abe N, Kato M, et al. Vildagliptin dose-dependently improves glycemic control in Japanese patients with type 2 diabetes mellitus. Diabetes Res Clin Pract 2009;83:233–40.
[27] Nonaka K, Kakikawa T, Sato A, et al. Efficacy and safety of sitagliptin monotherapy in Japanese patients with type 2 diabetes. Diabetes Res Clin Pract 2008;79:291–8.
[28] Fakhoury W, LeReun C, Wright D. A meta-analysis of placebo-controlled clinical trials assessing the efficacy and safety of incretin-based medications in patients with type 2 diabetes. Pharmacology 2010;86:44–57.
[29] Richter B, Bandeira-Echtler E, Bergerhoff K, Lerch CL. Dipeptidyl peptidase-4 (DPP-4) inhibitors for type 2 diabetes mellitus. Cochrane Database Syst Rev 2008;2:CD006739.
[30] DeFronzo RA, Stonehouse AH, Han J, Wintle ME. Relationship of baseline HbA1c and efficacy of current glucose-lowering therapies: a meta-analysis of randomized clinical trials. Diabetic Med 2010;27:309–17.
[31] Saurat JH, Stingl G, Dubertret L, et al. Efficacy and safety results from the randomized controlled comparative study of adalimumab vs methotrexate vs placebo in patients with psoriasis (CHAMPION). Br J Dermatol 2008;158:558–66.

[32] Menter A, Tyring SK, Gordon K, et al. Adalimumab therapy for moderate to severe psoriasis: a randomized, controlled phase III trial. J Am Acad Dermatol 2008;58:106–15.
[33] Leonardi CL, Powers JL, Matheson RT, et al. Etanercept as monotherapy in patients with psoriasis. N Engl J Med 2003;349:2014–22.
[34] Papp KA, Tyring S, Lahfa M, et al. A global phase III randomized controlled trial of etanercept in psoriasis: safety, efficacy, and effect of dose reduction. Br J Dermatol 2005;152:1304–12.
[35] Gordon KB, Langley RG, Leonardi C, et al. Clinical response to adalimumab treatment in patients with moderate to severe psoriasis: double-blind, randomized controlled trial and open-label extension study. J Am Acad Dermatol 2006;55:598–606.
[36] Signorovitch JE, Wu EQ, Betts KA, et al. Comparative efficacy of nilotinib and dasatinib in newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect comparison of randomized trials. Curr Med Res Opin 2011;27:1263–71.
[37] Cortes JE, Baccarani M, Guilhot F, et al. Phase III, randomized, open-label study of daily imatinib mesylate 400 mg versus 800 mg in patients with newly diagnosed, previously untreated chronic myeloid leukemia in chronic phase using molecular end points: tyrosine kinase inhibitor optimization and selectivity study. J Clin Oncol 2010; 28:424–30.
[38] Wei G, Rafiyath S, Liu D. First-line treatment for chronic myeloid leukemia: dasatinib, nilotinib, or imatinib. J Hematol Oncol 2010;3:47.
[39] Saglio G, Kim DW, Issaragrisil S, et al. Nilotinib versus imatinib for newly diagnosed chronic myeloid leukemia. N Engl J Med 2010;362: 2251–9.
[40] Kantarjian H, Shah NP, Hochhaus A, et al. Dasatinib versus imatinib in newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med 2010;362:2260–70.
[41] Biederman J, Melmed RD, Patel A, et al. A randomized, double-blind, placebo-controlled study of guanfacine extended release in children and adolescents with attention-deficit/hyperactivity disorder. Pediatrics 2008;121:e73–84.
[42] Sallee F, McGough J, Wigal T, et al. Guanfacine extended release in children and adolescents with attention-deficit/hyperactivity disorder: a placebo-controlled trial. J Am Acad Child Adolesc Psychiatry 2009;48: 155–65.
[43] Kelsey DK, Sumner CR, Casat CD, et al. Once-daily atomoxetine treatment for children with attention-deficit/hyperactivity disorder, including an assessment of evening and morning behavior: a double-blind, placebo-controlled trial. Pediatrics 2004;114:e1–8.
[44] Spencer T, Heiligenstein JH, Biederman J, et al. Results from 2 proof-of-concept, placebo-controlled studies of atomoxetine in children with attention-deficit/hyperactivity disorder. J Clin Psychiatry 2002;63: 1140–7.
[45] Michelson D, Allen AJ, Busner J, et al. Once-daily atomoxetine treatment for children and adolescents with attention deficit hyperactivity disorder: a randomized, placebo-controlled study. Am J Psychiatry 2002;159:1896–901.
[46] Michelson D, Faries D, Wernicke J, et al. Atomoxetine in the treatment of children and adolescents with attention-deficit/hyperactivity disorder: a randomized, placebo-controlled, dose-response study. Pediatrics 2001;108:E83.
[47] Intuniv package insert. Available from http://pi.shirecontent.com/PI/PDFs/Intuniv_USA_ENG.pdf. [Accessed May 18, 2012].
[48] Strattera package insert. Available from http://pi.lilly.com/us/strattera-pi.pdf. [Accessed May 18, 2012].
[49] Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683–91.
[50] Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005;9:1–134, iii–iv.